ARGONNE NATIONAL LABORATORY
9700 South Cass Avenue
Argonne, Illinois 60439

Rejoinder: Modeling an Augmented Lagrangian for Blackbox Constrained Optimization[1]

**Robert B. Gramacy, Genetha A. Gray, Sébastien Le Digabel,
Herbert K.H. Lee, Pritam Ranjan, Garth Wells, and Stefan M. Wild**

Mathematics and Computer Science Division

Preprint ANL/MCS-P5401-0915

September 2015

Rejoinder

Robert B. Gramacy[*]   Genetha A. Gray[†]   Sébastien Le Digabel[‡]

Herbert K.H. Lee[§]   Pritam Ranjan[¶]   Garth Wells[‖]   Stefan M. Wild[**]

September 22, 2015

We are grateful for the many insightful comments provided by the discussants. One team politely pointed out oversights in our literature review and the subsequent omission of a formidable comparator. Another made an important clarification about when a more aggressive variation (the so-called `NoMax`) would perform poorly. A third team offered enhancements to the framework, including a derivation of closed-form expressions and a more aggressive updating scheme; these enhancements were supported by an empirical study comparing new alternatives with old. The last team suggested hybridizing the statistical augmented Lagrangian (AL) method with modern stochastic search. Below we present our responses to these contributions and detail some improvements made to our own implementations in light of them. We conclude with some thoughts on statistical optimization using surrogate modeling and open-source software.

[*]Corresponding author: The University of Chicago Booth School of Business, Chicago, IL 60605; `rbgramacy@chicagobooth.edu`

[†]Intel Corporation, Folsom, CA 95630

[‡]GERAD and Département de Mathématiques et Génie Industriel, École Polytechnique de Montréal, Montréal, QC H3C 3A7, Canada

[§]Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA 95064

[¶]IIM Indore, Prabandh Shikhar Rau-Pithampur Road, Indore, M.P., India - 453556

[‖]Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, UK

[**]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439

1 Updates, initialization, and performance metrics

Picheny et al. made an important observation. When the AL parameters (λ, ρ) are updated aggressively, as is the case when the updates in Algorithm 1 are applied after every inner-loop step during which a candidate x^k is chosen, the average performance is improved. We can confirm that this is indeed the case, and some illustrations are provided along with a short comparison in Section 2. Our original updating scheme was designed conservatively, keeping in mind a statistical reinterpretation of the commonly applied AL updating rules (see, e.g., Nocedal and Wright, 2006). In our initial scheme, the inner loop is terminated—causing updates to occur—only after progress has plateaued, that is, when no change is seen in the expected improvement (EI) search under the current, fixed set of (λ, ρ) parameters. We did not experiment with these updates further because the initial scheme compared favorably with other methods.

Had we not overlooked an important comparator (Section 2), we likely would have focused more on the (λ, ρ) parameter updates. We appreciate now that the performance of the method during early iterations depends intimately on the frequency of the parameter updates and their initialization (especially for ρ). This new focus on early progress gave us a fresh perspective on hybrid statistical/mathematical programming strategies for blackbox constrained optimization, and how performance on that task is measured. Tracking the *best valid value* of the objective—a sensible metric in blackbox settings where feasible solutions are demanded and computational budgets limit evaluations—is well matched to many statistical optimizers, especially ones like EI, which involve little to no lookahead. EI, for example, has been shown to choose the next input as if it were its last (e.g., Bull, 2011). By contrast, AL methods actively search in invalid regions to ensure long-term progress. This property is at odds with our *best valid value* metric and is in fact amplified when using a global statistical strategy such as ours. A more aggressive update and initialization, when

paired with global response surface models and EI, turns out to be a better hybridization under that metric.

The initialization we now prefer involves choosing ρ^0 to balance the magnitude of objective function values $f(x)$ with the magnitude of the penalty term involving the squared constraint violations. In the AL given in (3), the factor $\frac{1}{2\rho}$ weights the latter term relative to the former, and we therefore choose ρ^0 so they are about the same nearby to the most promising objective and constraint values observed in an initial design $\{x_i, f(x_i), c_j(x_i)\}_{j=1}^{n_0}$. Specifically, let

$$\rho^0 = \frac{\min_{i=1,\ldots,n_0}\{\sum_{j=1}^m \max(c_j(x_i), 0)^2 : c_j(x_i) > 0 \text{ for some } j = 1, \ldots, m\}}{2\min_{i=1,\ldots,n_0}\{f(x_i) : c(x_i) \leq 0\}}.$$

The denominator is not defined if the initial design has no valid values (i.e., there is no x_i with $c_j(x_i) \leq 0$ for all j). In such cases, we use any of the values of f in the initial design (e.g., the median) in place of the undefined term in the denominator. Conversely, if the initial design has no invalid values and hence the numerator is not defined, we default to $\rho^0 = 1$.

This choice of initial penalty parameter ρ^0 ensures that the algorithm starts in a more neutral position in the sense of balancing the objective versus penalty through the constraints. Furthermore, this initialization has the added benefit of being invariant to scalar multiples of the objective and/or constraint functions (i.e., the effect of ρ^0 is unchanged for $\min\{\alpha f(x) : \beta c(x) \leq 0\}$ for any $\alpha, \beta > 0$). Whereas our experiments on the toy problem previously started with $\rho^0 = 1/2$, so that the initial weight was 1.0 on the quadratic penalty term, the new ρ^0 values (found via an initial uniform design of size $n_0 = 10$) are closer to $1/16$, giving a weight eight times greater. Thus this "more neutral" stance is more aggressive on this problem. The updating scheme of Picheny et al. ensures that it becomes even more aggressive as optimization trials evolve. Before examining the performance under this new

3

scheme, we consider a new comparator.

2 Overlooking an important benchmark

Chen and Welch politely pointed out an important oversight: Schonlau et al. (1998) provide a simple, EI-based scheme for handling multiple constraints that is easy to adapt to the known-objective case. Chen and Welch showed empirically that this method, which we call "EIC" for "expected improvement (under) constraints," compares favorably with our surrogate modeling AL hybrid scheme, as we originally presented it. To ensure that comparisons were just, we implemented our own version `optim.eic` (augmenting the original `optim.auglag`) in the `laGP` package for `R`; as the discussants suggested, this implementation was relatively straightforward. As can be seen in the left panel in Figure 1, the EIC comparator (solid red) reliably achieves the minimum after about 25 blackbox evaluations.[1]

These results, and those provided by Picheny et al., showed clear room for improvement in our hybrid AL implementation. Indeed, aggressive updates and more neutral initialization of the AL parameters (λ, ρ) lead to dramatic improvement on this toy example. The solid black line in the left panel of Figure 1 shows the average performance of this modified AL scheme, which is almost as good as EIC. EIC makes faster initial progress and may converge as many as five iterations earlier on average.

We have found that this behavior—slightly faster initial progress by EIC but ultimately both methods having similar convergence—is persistent across a wide range of synthetic test problems. For example, we created a harder, d-dimensional version of the toy problem. In this harder version, the second, ultimately nonbinding constraint which is the interior of a circle centered at the origin in the original $d = 2$ problem, is expanded to a d-ball through $c_2(x) = \sum_{i=1}^{d} x_i^2 - 3/2$. As the dimension d increases, the volume of the ball intersecting the

---

[1]This echoes the results provided by Chen and Welch; however, our search for the next point by EI does not utilize a simulated annealing.
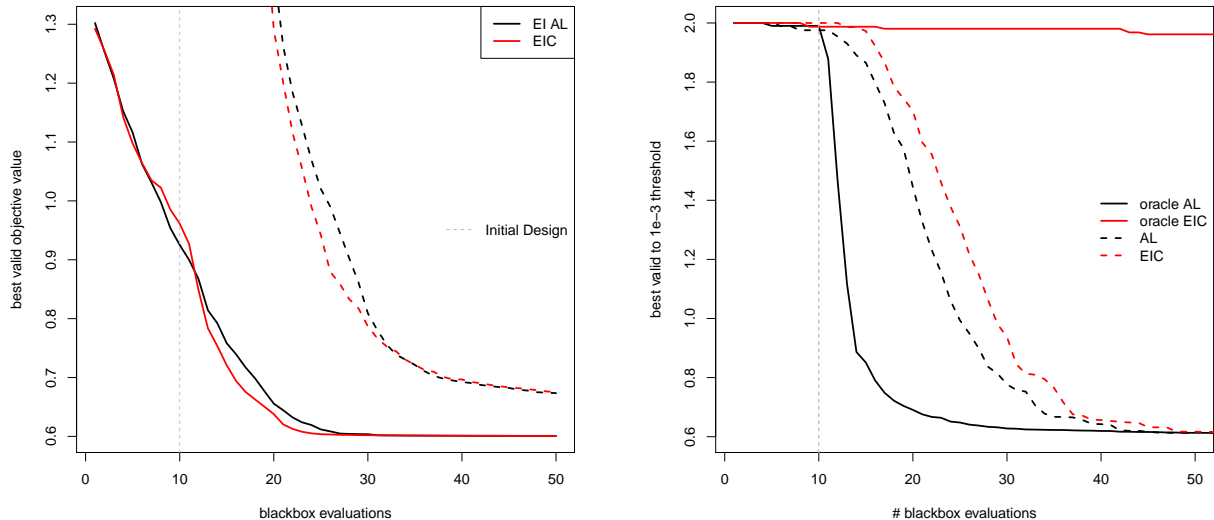
4

Figure 1: *Left:* a reworking of the comparison offered by Chen and Welch on a five-dimensional version of the toy problem; *right:* a variation on the toy problem with the second constraint replaced by the negative of the first constraint. The solid lines assume the constraint functions are known; the dashed lines model the constraint functions with GPs. All lines shown are average values calculated over 100 random initializations.

nonnegative orthant shrinks, thereby shrinking the valid region. The dashed lines in the left panel of Figure 1 show progress on the $d = 5$ case. The story is similar: slightly faster initial progress by EIC but then nearly identical performance afterwards.

Although EIC has many attractive features, the nature of how constraints enter into the selection criteria makes it prone to pathology when the valid region is very small. For a convincing illustration we consider a rather extreme case where the two constraints are the opposite of one another, inefficiently encoding an equality constraint. Adjusting our toy problem, we take $c_2(x) \equiv -c_1(x)$, so that the valid region has measure zero; that is, it is a submanifold along the sinusoidal curve traced out by the level set $c_1(x) = 0$, within the original two-dimensional space. We clarify that this is a well-posed problem in the framework targeted by the original manuscript,[2] and we acknowledge that more efficient ways exist for

---

[2]Technically, one would need to write $c_2(x) = -c_1(x) - \epsilon$ (with $\epsilon > 0$) for some of the theoretical results regarding constraint qualification to be justified.

handling equality constraints with the AL (e.g., as suggested by Picheny et al.). EIC struggles with this case because no part of the input space satisfies both constraints. If we suppose the constraints are known (i.e., rather than modeling and estimating probabilities for use in the EIC calculation, we use the true probabilities, 0 or 1, in the EIC expression), the performance is exceedingly poor, as shown by the solid red line in the left panel of Figure 1. By contrast, AL performs well in that setting (solid black line). Note that since the valid region has measure zero, we must relax the progress criteria. In this figure, for both AL and EIC comparators, we treat $c(x) \leq 10^{-3}$ as *approximately valid.*

AL does well because the criteria uses $c(x)$ values directly, rather than the $c(x) \leq 0$ values of EIC or probabilities thereof; EIC does poorly in part because mapping to probabilities (or Booleans) discards information. The dashed lines in the figure show a more realistic case, where the constraint functions are modeled with GPs. The distinction here is not as stark, but AL is still superior. EIC's paradoxically improved performance relative to a case where perfect information is available can be attributed to inefficiencies in the GP modeling code and conservative choices of priors on the parameters that govern the characteristic lengthscale and noise (nugget). Although we do not show these results, the closer those values are chosen to their (unknown) ideal settings, the worse the EIC method performs. The take-home message is that a better predictor for $c(x)$ can lead to worse performance by EIC.

3   `NoMax` and correlated outputs

We are grateful to Hare et al. for their comments on the applicability of an aggressive variation of the AL formulation that we dubbed `NoMax`. The discussants correctly pointed out—and offered convincing support with both theoretical and practical arguments—that this heuristic can lead to poor behavior by (incorrectly) making some nonbinding constraints

active. In particular, problems arise when the solver is initialized within the domain of attraction of a nonbinding constraint. However, we feel that the situation may be more nuanced in our particular context of global optimization (via AL) under expected improvement.

Indeed, as Hare et al. remarked, when objectives are linear and the solution lies on the boundary of one of the constraints, the `NoMax` heuristic works well. We add that, in our experience, it works well regardless of how the search is initialized. The results are not always superior, but we have not noticed their being pathologically bad. The explanation is that our EI search is global and is therefore less sensitive to initialization. One can, of course, engineer situations, for example by modifying the objective, where the pull of a nonbinding constraint is too great to be overcome by a global EI search. We remind the reader that we acknowledged the risk of `NoMax` in our original manuscript and do not advocate its use in general practice.

Hare et al. also correctly pointed out that correlated (or joint) modeling of all outputs may lead to improved response surface estimates and subsequently to improved EI calculations and faster convergence. For example, ideas along these lines are summarized in Chapter 6 of Santner et al. (2003). Unfortunately, one is rarely aware of how constraints may be correlated with one another or with the objective, especially when the simulator is a blackbox. In their discussion, Hare et al. offered an example comparing independent models with correlated models, and the results were mixed. Our own experience is different. When assuming one of the standard multi-output modeling apparatuses—for example, *co-kriging* in the style of Mardia and Goodall (1993)—we find that the tight coupling of correlated outputs leads to poorer prediction compared with otherwise independent modeling of the spatial fields. This happens even when *known* correlation is present between the fields. The reason is that the assumption of a common/shared lengthscale (and global scale), as typically deployed when *co-kriging*, is rarely appropriate.

One important exception may be when the objective is known to be anti-correlated with

some constraint(s). This is a typical situation, one exemplified by both our examples, and one that may not require peeking into the blackbox to confirm. One can even argue that constrained optimization problems are hard precisely because at least one of the constraints typically operates in opposition to the objective function, in other words, that the objective function is lowest in a region where at least one constraint is not met. In such a setting, we have found that directly acknowledging negative correlation in the response surface model(s) improves results (see, e.g., Pourmohamad and Lee, 2015). Indeed, even more flexible modeling can allow fitting of negative correlations for active constraints and no correlations for constraints that do not interact with the minimum.

4   Taxonomies, annealing, and final thoughts

In their discussion, Picheny et al. presented a characterization of simulation-based constrained optimization problems. We emphatically agree that solution approaches may fundamentally differ depending on the specific nature of the constraints; such characterizations are thus critical for algorithm development and benchmarking. The characterization of Picheny et al. is based on the (relative) computational expense of the constraint and objective functions. A more general taxonomy of simulation-based constraints is the QRAK taxonomy of Le Digabel and Wild (2015). In addition to distinguishing between *a priori* and *simulation-based* constraints (which respectively can be coarsely viewed as cheap, algebraically available and expensive, blackbox constraints), the QRAK taxonomy captures information about the constraint functions that could be useful for statistical modeling purposes.

In particular, two other distinctions in QRAK are whether a constraint output is *quantifiable* or *nonquantifiable* (i.e., nonordinal) and whether a constraint must be satisfied in order to get meaningful output from the simulation outputs (*unrelaxable*, as opposed to the

complementary *relaxable* case). As a specific example, each of Picheny et al.'s first three cases likely has an implicit assumption that the constraint functions are quantifiable and relaxable; this is an assumption that we also make for our AL method. The final distinction in QRAK is whether a simulation-based constraint is *known* or *hidden*, the latter being related to Picheny et al.'s fourth case when the simulation crashes and no further output or flags indicate the reason for the crash. We agree that statistical methods can play a role in addressing problems with various combinations of each constraint type.

We thank Cheng and Liang for their discussion of simulated annealing methods for optimization. Stochastic approximation annealing (SAAn) can significantly improve the convergence times of the algorithm; use of population methods can further improve efficiency. Although the papers cited address only unconstrained optimization, Cheng and Liang suggested that using this approach on the AL function could achieve the best of both worlds. We would be interested in seeing the results of this hybrid approach. A complication, however, is that the AL is not a fixed function in x as is typically assumed of an objective function. Instead, the AL depends on parameters λ and ρ that are updated during the optimization. We thus wonder how well SAAn would adapt to such a moving target. Also unclear is whether this hybrid algorithm would retain convergence guarantees. Should SAAn not be sufficiently adaptable for a moving target, one could apply it in the context of more traditional penalized approaches, where a fixed penalty parameter attempts to drive the optimization into the region of feasibility.

We remind the reader that the AL framework, described in the original paper and embellished here thanks to the thoughtful discussions, has been fully implemented in the `laGP` package for `R`. To address a comment from Chen and Welch, we note that the software provides an option allowing the objective to be modeled with a GP, even though our original article (and this rejoinder) ignored that case for simplicity. We also understand from Picheny et al. that AL has been implemented in `DiceOptim` (Ginsbourger et al., 2015); this is excel-

lent news. In our view, open-source software is sorely lacking for surrogate-modeling-based approaches to optimization, constrained or otherwise. The *Journal of Statistical Software* recently published a special issue on optimization in R (volume 60, 2014), but it is troubling that no article therein highlights a *statistical* methodology applied to optimization. Obviously, such codes exist; but the information is not spreading as rapidly as we would hope. These are powerful techniques that are relatively straightforward to apply—especially ones like EIC—given mature GP response surface modeling libraries. We our hope that these and future open-source projects become more widely recognized in the literature.

Acknowledgments

References

Bull, A. D. (2011). "Convergence Rates of Efficient Global Optimization Algorithms." *J. of Machine Learning Research*, 12, 2879–2904.

Ginsbourger, D., Picheny, V., Roustant, O., with contributions by C. Chevalier, Marmin, S., and Wagner, T. (2015). DiceOptim*: Kriging-Based Optimization for Computer Experiments*. R package version 1.5.

Le Digabel, S. and Wild, S. M. (2015). "A Taxonomy of Constraints for Simulation-Based

Optimization." Preprint ANL/MCS-P5350-0515, Argonne National Laboratory, Mathematics and Computer Science Division.

Mardia, K. V. and Goodall, C. R. (1993). *Spatial-Temporal Analysis of Multivariate Environmental Monitoring Data*, vol. 6, 347–385. North Holland.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. 2nd ed. Springer.

Pourmohamad, T. and Lee, H. (2015). "Multivariate Stochastic Process Models for Correlated Responses of Mixed Type." Tech. Rep. 15-08, University of California, Santa Cruz, School of Engineering.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. New York: Springer-Verlag.

Schonlau, M., Jones, D. R., and Welch, W. J. (1998). "Global versus Local Search in Constrained Optimization of Computer Models." In *New Developments and Applications in Experimental Design*, vol. 34, 11–25. Institute of Mathematical Statistics.